scientific reports

Check for updates

OPEN

Evolution of respiratory syncytial virus genotype BA in Kilifi, Kenya, 15 years on

Everlyn Kamau[1,4]✉, James R. Otieno[1,5], Clement S. Lewa[1], Anthony Mwema[1], Nickson Murunga[1], D. James Nokes[1,2] & Charles N. Agoti[1,3]

Respiratory syncytial virus (RSV) is recognised as a leading cause of severe acute respiratory disease and deaths among infants and vulnerable adults. Clinical RSV isolates can be divided into several known genotypes. RSV genotype BA, characterised by a 60-nucleotide duplication in the G glycoprotein gene, emerged in 1999 and quickly disseminated globally replacing other RSV group B genotypes. Continual molecular epidemiology is critical to understand the evolutionary processes maintaining the success of the BA viruses. We analysed 735 G gene sequences from samples collected from paediatric patients in Kilifi, Kenya, between 2003 and 2017. The virus population comprised of several genetically distinct variants ($n = 56$) co-circulating within and between epidemics. In addition, there was consistent seasonal fluctuations in relative genetic diversity. Amino acid changes increasingly accumulated over the surveillance period including two residues (N178S and Q180R) that mapped to monoclonal antibody 2D10 epitopes, as well as addition of putative N-glycosylation sequons. Further, switching and toggling of amino acids within and between epidemics was observed. On a global phylogeny, the BA viruses from different countries form geographically isolated clusters suggesting substantial localized variants. This study offers insights into longitudinal population dynamics of a globally endemic RSV genotype within a discrete location.

Respiratory syncytial virus (RSV) is a leading cause of severe lower respiratory tract infections, with an estimated 3.2 million annual hospitalizations and approximately 60,000 deaths globally, primarily in children younger than 1-year-old[1–3]. Among adults, RSV produces a wide range of clinical symptoms including upper respiratory tract infections, severe lower respiratory tract infections, and exacerbations of underlying disease[4,5]. RSV epidemics exhibit clear patterns of seasonality and repeat infections are common throughout life[6]. Despite its global health impact, effective RSV prophylactics are limited, but multiple vaccine candidates and monoclonal antibodies are in different stages of development and licensure[7,8].

RSV genome, a negative-sense single-stranded RNA molecule, encodes 11 proteins including the fusion (F) and attachment (G) glycoproteins, which mediate virus binding and entry into host cells and are targets for neutralizing antibodies[9]. Two groups, RSVA and RSVB, defined based on antigenic and genetic variability within the glycoproteins co-circulate worldwide, causing seasonal epidemics[10,11]. Several RSV genotypes, 10 for RSVA and 13 for RSVB, have been identified to date[12,13]. In 1999, the RSVB genotype BA emerged with a 60-nucleotide duplication in the G gene and disseminated globally, replacing all other RSVB genotypes[14,15]. It is likely the additional residues of the 60-nucleotide duplication provide a selective advantage and modified the antigenic characteristics of the G protein, allowing escape from antibody neutralization, as well as enhancing fitness[15].

RSV longitudinal surveillance studies in Kilifi, coastal Kenya, have revealed a high disease burden particularly in infancy and early childhood[4,16–23]. Studying the genetic and antigenic evolution of local strains is crucial for understanding how annual RSV epidemics are maintained locally. This may help with designing comprehensive infection reduction/control measures and rationalize respiratory virus epidemic response policies[24,25]. Studies of the origins of RSV seed strains for the local regular epidemics, hubs of infection transmission, spread patterns, viral evolution and reinfection patterns would lead to better epidemic management[26] as recently demonstrated for emerging viruses like Ebola, Zika, SARS-CoV, SARS-CoV-2, Influenza A/H1N1/09)[27,28].

[1]Epidemiology and Demography Department, Kenya Medical Research Institute (KEMRI) – Wellcome Trust Research Programme, Kilifi, Kenya. [2]School of Life Sciences and Zeeman Institute (SBIDER), University of Warwick, Coventry, UK. [3]School of Health and Human Sciences, Pwani University, Kilifi, Kenya. [4]Present address: Nuffield Department of Medicine, University of Oxford, Oxford, UK. [5]Present address: Fogarty International Center, NIH, Bethesda, MD, USA. ✉email: everlyn.kamau@ndm.ox.ac.uk

RSV epidemic	RSV positive samples	Number (%) of RSV-B samples[a]	RSVB genotype BA sequences	Number (%) of RSVA samples[a]
2002/3	89	35 (39.3)	4	42 (47.2)
2003/4	114	16 (14)	6	68 (59.6)
2004/5	183	**119 (65)**	152	46 (25.1)
2005/6	239	9 (3.8)	9	224 (93.7)
2006/7	195	22 (11.3)	22	153 (78.5)
2007/8	256	**197 (77)**	141	33 (12.9)
2008/9	208	41 (19.7)	14	154 (74)
2009/10	259	95 (36.7)	58	120 (46.3)
2010/11	279	21 (7.5)	17	241 (86.4)
2011/12	161	**102 (63.4)**	81	55 (34.2)
2012/13	152	27 (17.8)	14	110 (72.4)
2013/14	117	42 (35.9)	42	75 (64.1)
2014/15	202	69 (34.2)	53	133 (65.8)
2015/16	149	83 (55.7)	72	66 (44.3)
2016/17	76	**53 (69.7)**	50	23 (30.3)

Table 1. Number of RSVB positive samples and genotype BA sequences by epidemic in Kilifi, 2002 to 2017. Bold values indicate that the proportion of RSV-B samples was greater than RSV-A samples. [a]Percentage of RSV positive samples.

The RSVB genotype BA was first detected in Kilifi in early 2003 and became dominant from 2004 onwards except for sporadic detection of the genotype SAB4 between the years 2011 and 2013[24]. We analyzed the G glycoprotein of RSVB genotype BA strains collected over 15 successive RSV epidemics in Kilifi. We show that locally, the genotype is characterized by discrete temporal genetic clustering and sequential variant replacement between and sometimes within epidemics. This work extends previous analyses[18,24] by a further six years of new data and augments our knowledge of the evolutionary trajectory and adaptation of a single RSV genotype in a single local setting when observed longitudinally.

Results

Over 15 RSV epidemics (2002 to 2017), a total of 903/12203 (7.4%) respiratory samples tested positive for RSVB. G gene sequencing was attempted for all RSVB positive samples and was successful for 857 (95%) RSVB positive samples. The proportion of RSVA and RSVB samples in each epidemic ranged from 3.8 to 77.0% (Table 1). RSVB dominated the 2004/5, 2007/8, 2011/12 and 2016/17 epidemics and co-dominated with RSVA in the 2009/10 and 2014/15 epidemics (Table 1). The annual distribution of RSV cases by age from 2002 to 2017 are shown in Supplementary Table 1. Although the hospital-based RSV surveillance in Kilifi was established in 2002, the genotype BA was not detected until January 2003. From the RSVB positive samples, RSVB BA viruses were confirmed during sequence assembly and analysis by the presence of a 60-nucleotide duplication in the C-terminal of the G gene. Overall, a total of 735 genotype BA G gene sequences spanning nucleotides 214 to 981 of the RSVB strain B1 (DQ227363) were obtained.

Genetic diversity of the genotype BA. The sequenced G gene region showed 3% nucleotide divergence (overall mean *P* distance) over the entire period and varied by a maximum of 38 nucleotides within an epidemic. Genetic divergence increased proportionally with time, Fig. 1A, indicating clock-like evolution. Molecular clock phylogenies showed a well-ordered diversification of the genotype BA since its introduction in Kilifi (Fig. 1B). Viruses from the same epidemic formed multiple phylogenetic clusters and often, sequences from the preceding epidemic were often positioned at the basal nodes of those in the successive epidemic, suggesting sequential virus circulation and persistence between epidemics. A total of 56 distinct variants were identified, some were singletons suggesting either low level transmission or unsampled genetic diversity. The 2007/8 and 2009/10 epidemics were the most heterogeneous, with 12 and 11 variants, respectively. Up to four variants circulated over multiple epidemics (Fig. 1C), and some were observed in non-consecutive epidemics likely indicating re-introduction of variants that had been previously observed locally.

Evolutionary history of the genotype BA in Kilifi. Our previous study reported the number of virus introductions into Kilifi from 2002/3 to 2011/12 epidemics[24]. To place the Kilifi genotype BA viruses in the context of global RSVB diversity, we analyzed the 2012/13 to 2016/17 Kilifi sequences ($n = 233$) and 583 global sequences sampled from 16 countries between 2012 and 2017. In the current study, distinct geographical clusters were evident with viruses clustering primarily by country of origin. The 2012/13 to 2016/17 RSVB epidemics in Kilifi were seeded by at least 15 virus introductions (Fig. 2). For many countries, viruses circulating during the same year were not placed into single monophyletic groups but in multiple clusters of assorted sizes (Fig. 2), indicative of multiple virus entries followed by local spread and genetic expansion. Between 2012 and 2017,

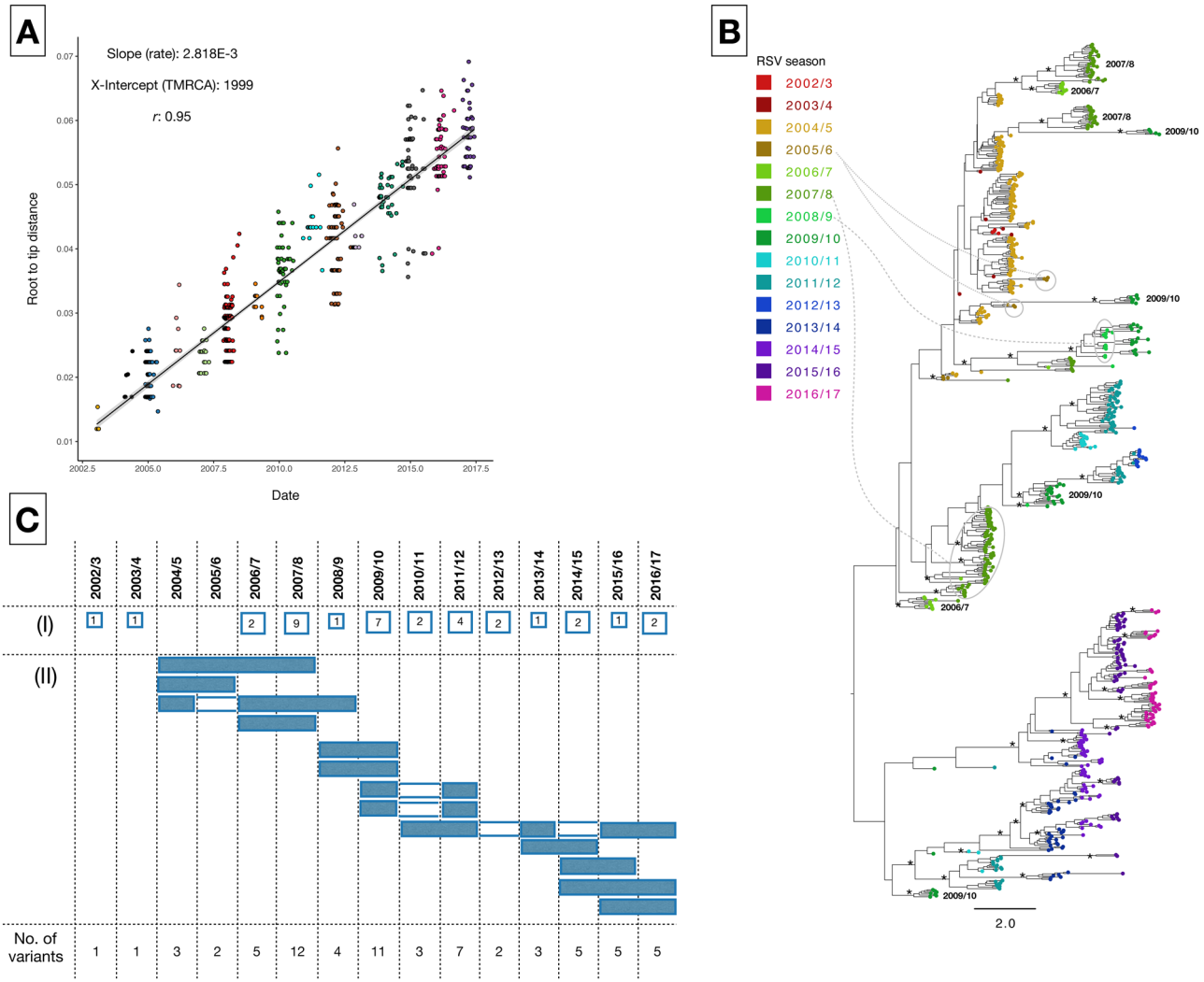

Figure 1. (**A**) Correlation plot of root-to-tip genetic distance against sampling date for a phylogeny estimated from RSVB genotype BA 735 G gene sequences sampled from Kilifi, Kenya. The estimated correlation coefficient and R^2 values were 0.95 and 0.9, respectively. (**B**) Maximum clade credibility (MCC) tree inferred for 735 G gene sequences (756 nucleotides) from Kilifi, with tip labels colored by RSV epidemic. Node support is indicated by (*) for posterior probabilities > 0.9. (**C**) Temporal occurrence of the 56 RSVB genetic variants (rows) identified in Kilifi between 2003 and 2017. A variant was defined as a virus or group of viruses with ≥ 4 nucleotide differences in the sequenced G region (see "Methods" section). The number of variants circulating only in a single epidemic are shown in (I) and variants that circulated in more than one epidemic are indicated by filled rectangles in (II).

several other variants circulated outside Kilifi suggesting that contemporaneous RSV epidemics are as a result of different localized variants and less likely sequential transmission between countries.

Virus population dynamics. We used Bayesian skyride analysis to estimate temporal changes in relative genetic diversity (Fig. 3A). Elevation in relative genetic diversity was correlated with increased RSVB activity and appearance of new viral populations in the 2004/5, 2007/8, and 2016/17 epidemics (Fig. 3B). Demographic expansions were interspersed by constrictions in the effective population size for example in 2006, 2009/10, and 2014/15, characteristic of bottleneck effects and variant replacement between epidemics[15,29]. It was previously suggested that the expansion in the effective population in 2005 coincided with the predominance and rapid dissemination of the genotype BA, relative to the rest of the group B viruses[15]. The increase in relative genetic diversity between 2006 and 2008 could also be related to the introduction and dissemination of viruses with novel mutations in 2007/8 as described below.

Amino acid changes. Numerous amino acid (aa) changes occurred with variable frequencies over the 15 epidemics, and commonly these changes involved reversal or 'toggling' between codons within or between epidemics. Codon positions exhibiting frequent (> 10% of sequences) reversals are marked in red in Fig. 4. The 2002/3 to 2006/7 epidemics had few changes, while the 2007/8 epidemic saw increased RSVB activity and

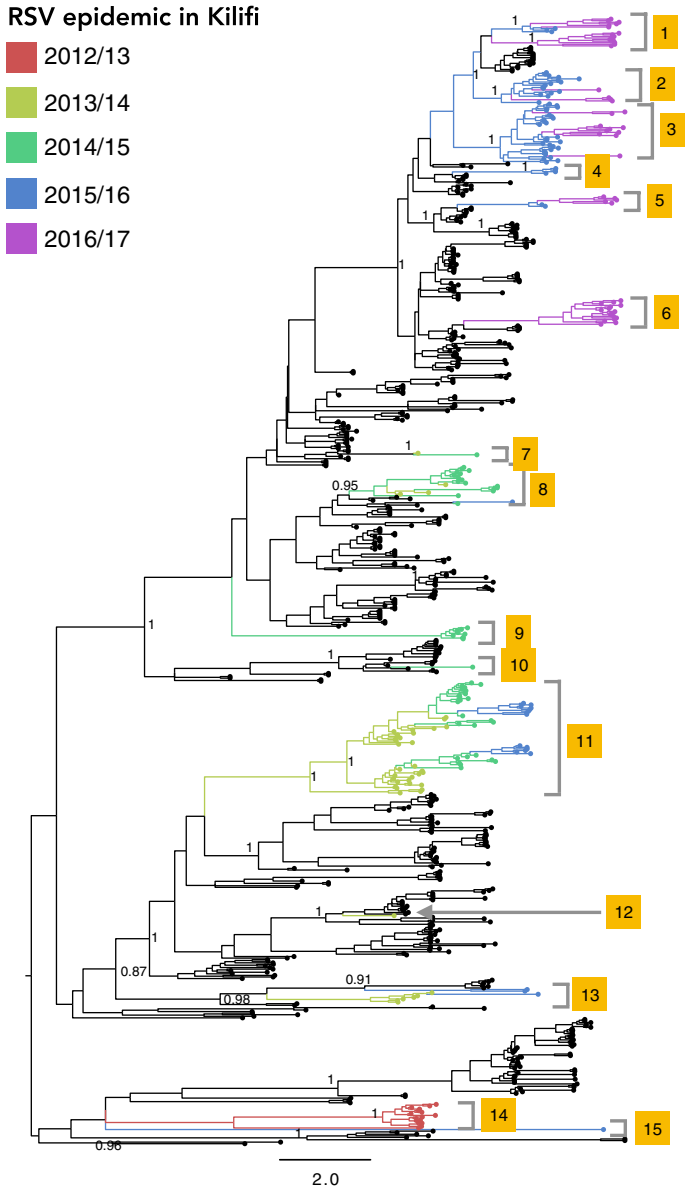

Figure 2. Maximum likelihood phylogeny estimated from 716 genotype BA G gene sequences sampled globally between 2012 and 2017. Discrete virus introductions in Kilifi are marked on the phylogeny. Terminal branches with sequences from Kilifi are coloured by RSV epidemic. Terminal branches with black circles are indicate sequences from locations outside Kilifi. Clade posterior probabilities are shown for selected nodes (>0.75).

appearance of several new amino acid changes. Additional changes emerged in every epidemic after 2007/8 particularly in RSV seasons with high RSVB incidence. Several positions had more than one amino acid change, including T108I/A, R137K/T, P206L/S, E241G/K, S257L/P, K258I/Q, S269A/P/F, E304K/D/N and K314G/R.

Occasionally, new and possibly beneficial variants emerged and became fixed in the viral population (red dotted borders, Fig. 4). For example, the I199T amino acid substitution was introduced in 2011/12 and Threonine (T) became fixed at position 199 in the subsequent epidemics, while the T108A, T254I, I280T and R314K substitutions became fixed in the 2015/16 and 2016/17 epidemics. We observed amino acid switch from 'minority' to 'majority' variant: when a 'minor' amino acid variant (<30% frequency) occurred in the majority (>70%) of viruses in the successive epidemic(s) or re-appeared in the population after absence in preceding epidemic(s) (blue dotted borders, Fig. 4). For instance, Leucine (position 218) reappeared in 2013/14 and became dominant through to 2016/17 epidemic. This is also evident for Histidine at position 286. The emergence of viruses with Q313Stop mutation occurred in 2007/8 akin to a previous report[15], then disappeared in 2008/9, reappeared in 2009/10, and disappeared again until 2013/14. This codon has been proposed to be under positive selection[30].

We also examined variability at the 60-nucleotide analogous (aa 241–259) and duplicated segments (aa 260–279) (Fig. 5). Relative to oldest BA sequences (2002/3) in Kilifi, the 2003/4 to 2005/6 viruses were nearly

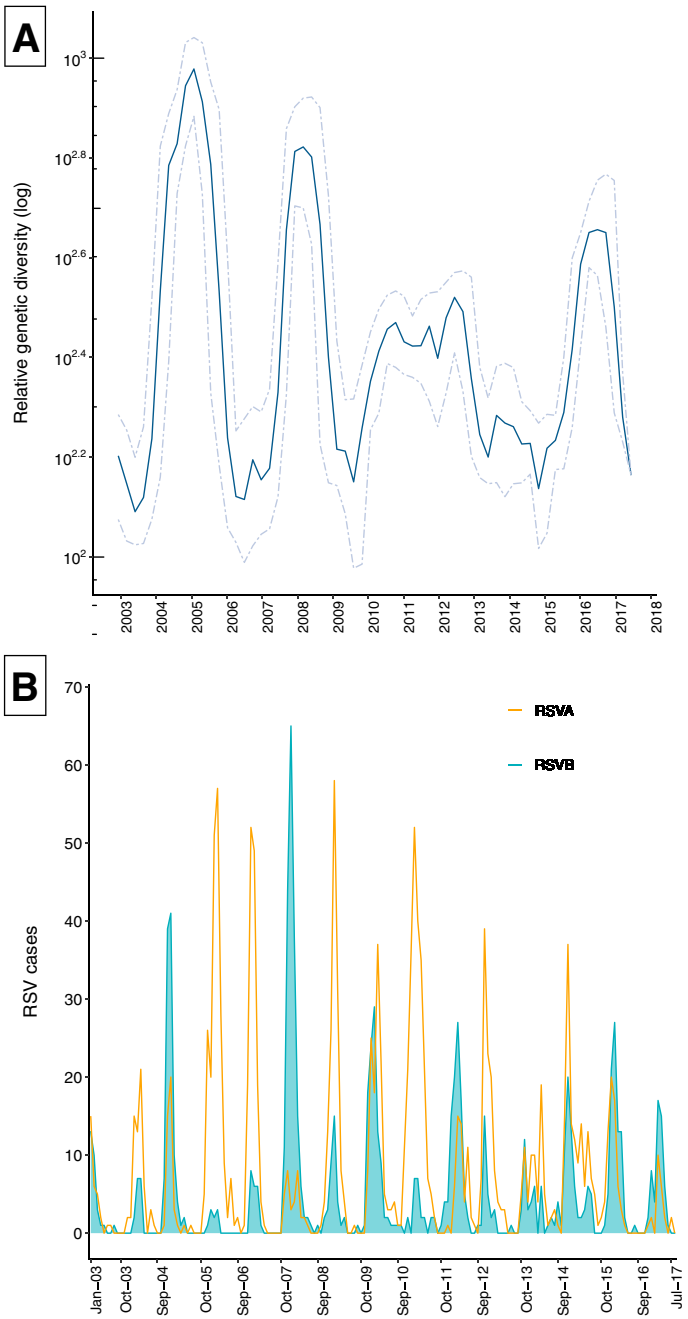

Figure 3. (**A**) Relative genetic diversity estimated using the Gaussian Markov Random Field (GMRF) Bayesian skyride coalescent model. Solid lines represent mean relative genetic diversity while the dotted lines indicate the 95% HPD intervals. (**B**) Epidemic patterns of RSV antigenic groups (RSVA and RSVB) in Kilifi, Kenya, from 2002 to 2017.

identical in both segments. Amino acid changes gradually accumulated in the duplicated segment from the 2007/8 epidemic.

Sequence variations at the central conserved domain. Between the two mucin-like regions of RSV G is an un-glycosylated central conserved domain (CCD) (positions ~155 to 206), with a stretch of 13 aa (164–176) that is strictly maintained in all strains, and four invariant cysteines (residues 173, 176, 182 and 186) with a 1–4, 2–3 disulfide topology forming a cysteine noose that are also highly conserved[31,32]. The third and fourth cysteines (Cys^{182}–Cys^{186}) form a CX3C motif poised for interaction with CX3CR1, mediating RSV attachment during infection[31]. Human broadly neutralising monoclonal antibodies (mAb) have been reported to bind in the CCD[33–35].

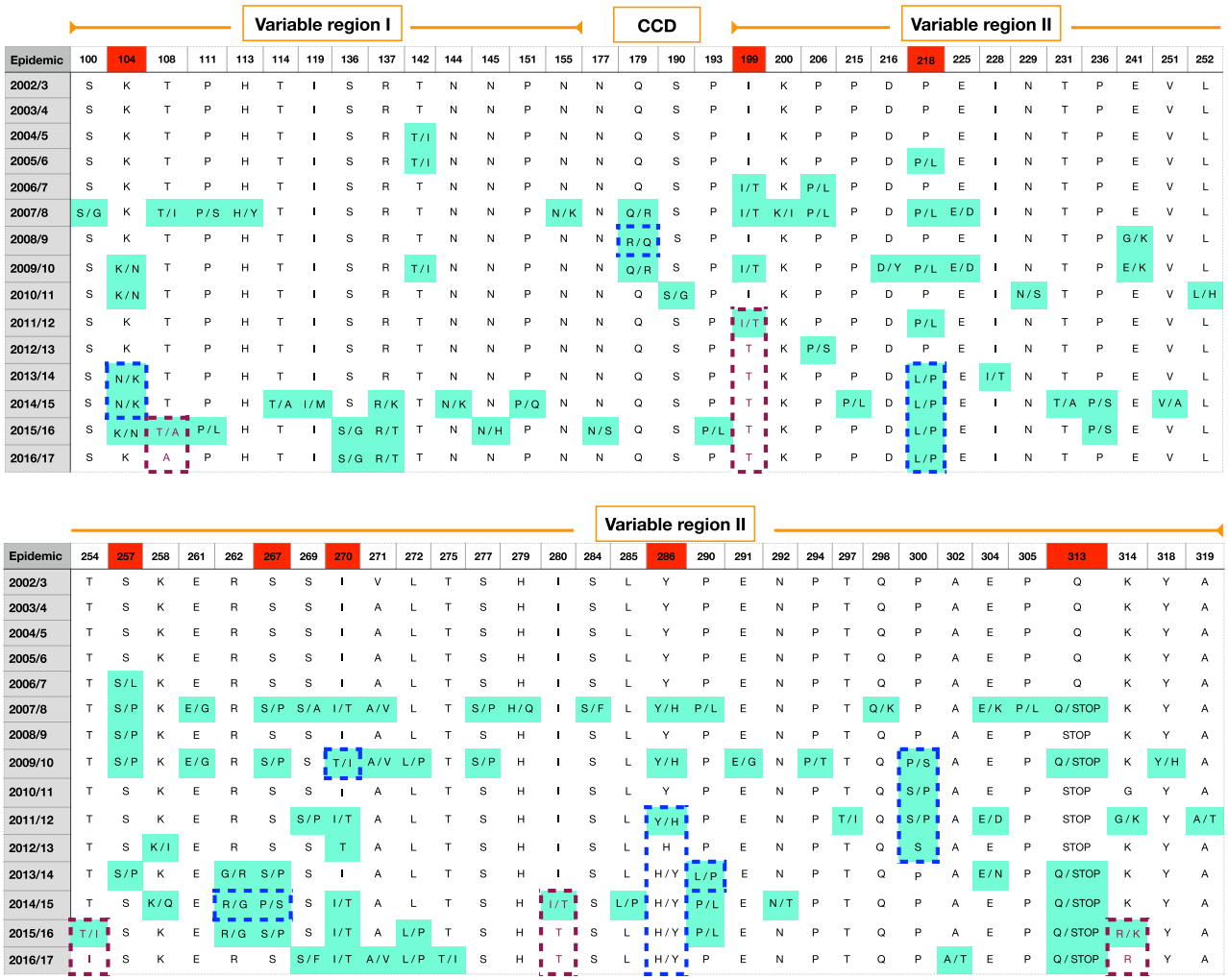

Figure 4. Amino acid differences in the G protein relative to majority consensus residue for each position indicated. Shown are the differences or amino acid variants detected in more than 10% of the viruses. Red dotted borders indicate new and possibly beneficial variants that emerged and became fixed in the viral population. Blue dotted borders indicate amino acid switch from 'minority' to 'majority' variant.

We found two amino acid substitutions (N178S and Q180R) located within the cysteine noose. The N178S substitution was detected in 14 sequences in the 2014/15 epidemic, while the Q180R mutation appeared in 2007/8 and persisted through to 2009/10. Importantly, Arginine at position 180 was detected or fixed in all the strains sampled in 2008/9. These aa changes are positioned within the antigenic site γ2, the conformational epitope (residues 177–188) of the high affinity broadly neutralising human mAb 2D10[33]. Structurally, the two residues are oriented outward, residue 178 is visible in a loop connecting two alpha-helices and residue 180 occurs in an alpha helix. The two substitutions are relatively conservative (polar, hydrophilic) and unlikely to have a substantial effect on G structure or function but could be molecular adaptations to natural selection pressure.

N-linked glycosylation. We identified N-linked glycosylation sequons at 14 residues (81, 86, 117, 134, 144, 230, 243, 263, 265, 273, 293, 296, 305 and 310) occurring at varied frequencies (Table 2). The number of N-glycosylation sites increased continuously over the epidemics. The minimum number of possible N-glycosylation sites per sequence was three, and the maximum was four. Two putative N-glycosylation sites (296 and 310) were dominant in the majority (93.2% and 98.1%, respectively) of the sequenced viruses. Amino acid substitutions (N296K/D and S297P), and T312N/A/I, occurring between 2010/11 and 2013/14 resulted in loss of potential N-glycosylation sequons at position 296 and 310, respectively.

Notably, two new putative N-glycosylation sequons (positions 81 and 86) emerged in 2014/15 and persisted in all the viruses sampled the succeeding epidemics. Changes in glycosylation state can be a source of antigenic novelty conferring selective advantage by 'masking' epitopes from antibody recognition. It was observed that addition of glycosylation sites to hemagglutinin in Influenza A can reduce or abolish binding by monoclonal antibodies or human antisera[36].

Analogous region: aa positions 240-259

	240	241	242	243	244	245	246	247	248	249	250	251	252	253	254	255	256	257	258	259
2002/3	T	E	R	D	T	S	T	P	Q	S	T	V	L	D	T	T	T	S	K	H
2003/4																				
2004/5																				
2005/6																				
2006/7																		S / L		
2007/8																		S / P		
2008/9		G / K																		
2009/10		E / K		D / N								V / A						S / P		
2010/11													H							
2011/12							I													
2012/13																			I	
2013/14																		S / P		
2014/15							T / A					V / A							K / Q	
2015/16															T / I					
2016/17															I					

Duplicated region: aa positions 260-279

	260	261	262	263	264	265	266	267	268	269	270	271	272	273	274	275	276	277	278	279
2002/3	T	E	R	D	T	S	T	S	Q	S	I	V	L	D	T	T	T	S	K	H
2003/4												A								
2004/5												A								
2005/6												A								
2006/7												A								
2007/8		E / G						S / P		S / A	I / T	A / V						S / P		H / Q
2008/9												A								
2009/10		E / G						S / P			I / T	V / A	L / P					S / P		
2010/11												A								
2011/12								S / P		S / P	I / T	A								
2012/13											T	A								
2013/14			R / G					S / P				A								
2014/15			R / G					S / P			I / T	A	L / P							
2015/16			R / G					S / P			I / T	A / V	L / P							
2016/17								S / P		S / F	I / T	A / V	L / P			T / I				

Figure 5. Amino acid changes in the 60-nt (20-aa) duplication region relative to the earliest RSVB genotype BA sequences in the Kilifi dataset.

Selection analyses. Evidence for pervasive and episodic selection was found in 13 codon positions: 144, 151, 159, 223, 227, 270, 281, 287, 289, 290, 305, 306, 313. Eight of the 13 positions (underlined) were identified by more than one method. The MEME method found positions 281, 305 and 313 under episodic adaptive selection (occurring heterogeneously across branches), while the FUBAR and FEL methods identified pervasive selection in positions 159 and 306, respectively. The SLAC method estimated the overall mean nonsynonymous/synonymous (d_N/d_S) substitution rate ratio as 0.56. Generally, the analyzed G region is under purifying/negative selection ($d_N/d_S < 1$), and only a very restricted number of codon positions have evolved at a higher non-synonymous substitution rate (> 1, in Supplementary Figure S1).

Discussion

This study documents the natural history of the RSVB genotype BA viruses in Kilifi, Kenya, over 15 epidemics. Our analysis found substantial genetic diversity exhibited by multiple variants co-circulating within and between epidemics. Shifts in the predominating strains have been suggested to confer a selective advantage by variations

Pattern	Codon position												Epidemics observed
	81	86	127	134	230	243	263	265	273	293	296	310	
1	–	–	–	–	–	–	–	–	–	–	+	+	2002/3, 2003/4, 2004/5, 2005/6, 2006/7, 2007/8, 2008/9, 2009/10, 2010/11, 2011/12, 2012/13, 2013/14, 2014/15
2	–	–	–	–	–	–	–	–	–	+	+	+	2009/10
3	–	–	–	–	–	+	–	–	–	–	+	+	2009/10
4	–	–	+	–	–	–	–	–	–	–	+	+	2007/8
5	–	–	–	–	+	–	–	–	–	–	+	+	2014/15
6	–	–	–	+	–	–	–	–	–	–	+	+	2006/7
7	–	–	–	–	–	–	–	+	–	–	+	+	2004/5
8	–	–	–	–	–	–	+	–	–	–	+	+	2009/10
9	–	–	–	–	–	–	–	–	+	–	+	+	2004/5
10	+	+	–	–	–	–	–	–	–	–	+	+	2014/15, 2015/16,2016/17
Frequency (%)	17.5	17.5	0.4	0.01	0.01	0.7	0.01	0.01	0.01	1	93.7	98.6	

Table 2. N-linked glycosylation sites and patterns and their occurrence in the RSV-B epidemics in Kilifi, Kenya.

in strain- or clade-specific immunity that favors circulation of heterologous variants[37,38]. Virus sequences from successive epidemics clustered together, suggesting variant persistence between RSV seasons. As previously implied[39], it is possible that genetically close viruses are re-imported each year from unsampled locations rather than persisting locally.

The demographic expansions around 2005 and 2008 denote periods of optimal adaptation for virus replication and widespread dissemination, congruent with proliferation of new lineages less susceptible to immune responses[40], and were previously described in relation to the global spread of the BA-IV lineage viruses[15]. Fluctuation in relative genetic diversity around 2003 and early 2007 is characteristic of bottleneck effects or 'viral eclipse phases' preceding population expansion, probably imposed by selection pressure or as a result of decline of circulating variants[40].

The 60-nucleotide duplication region showed further accumulation of nucleotide substitutions further than previously reported[15], which likely points to enhanced fitness and molecular adaptation providing evolutionary advantage of the BA genotype[41] compared to other RSVB genotypes. Accumulation of nucleotide changes in RSV G gene was also attributed to antibody selection to some extent, as well as selective constraints other than immune selection, for instance, high mutation rate, protein malleability and bottleneck effects[42].

We observed amino acid reversion patterns in the Kilifi genotype BA viruses. Evolutionary reversals provide a means of escape by generating antigenic novelty and reflect fluctuating dynamics in immune responses, in which cross-immunity wanes, thereby controlling pathogen replication and facilitating virus transmission[30]. Escape mutations within or flanking functionally conserved epitopes come under selective pressure to revert to the wild type in hosts that do not mount an immune response against the epitope[43]. Amino acid reversions could also be a necessary consequence of a limited set of possible replacements at epitopes, a constraint on the repertoire of functionally viable amino acids[30].

The Arg-180 substitution may have altered the positive charge of the G protein cysteine noose. The role of electrostatic charge in pathogenesis has been appreciated for HIV rapid progression to AIDS by the presence of charged residues at two amino acid positions on the gp120 subunit[44]. In Influenza A, several immune escape mutants showed an increased positive charge in HA and virions with a higher net positive charge in HA may promote favourable electrostatic interactions with target cells since host cell receptors possess a strong negative charge[45]. A change in surface charge (Glu to Lys) of the envelope glycoprotein was also implicated in the emergence of enzootic and epizootic viral phenotypes[46]. In addition, a potentially significant element of the immune escape repertoire is the selection of 'adsorptive mutants', showing enhanced binding to target cells[47]. An important aspect of receptor-binding avidity is that of electrostatic charge[45]. The Arginine replacement at position 180 may have resulted in increased net positive surface charge at the cysteine noose consequently promoting electrostatic interactions with target host cells. Future therapeutic design efforts will need to assess for possible loss of human G-directed mAbs cross-reactivity to epitopes containing the N178S and Q180R mutations.

Both pervasive positive selection and episodic diversifying selection were detected in amino acid positions within the sequenced G gene segment, six of which (227, 270, 287, 305, 306 and 313) were in agreement with previous studies[30,48]. Similarly, four putative N-glycan binding sites reported in this study (86, 144, 296, and 310) were previously reported for RSV G[49]. Amino acid position 144 was also identified in our study as being under positive selection by at least two methods.

In summary, our analyses show that the genotype BA dynamics in Kilifi have been marked by multiple co-circulating variants even within an epidemic, which are serially replaced over time perhaps by virus importations or evolution in situ. There has been gradual diversification specially in the C-terminal of the G protein possibly due to antibody-driven selection or functional constraints. In later epidemics (from 2007/8) the genotype was characterized by marked evolutionary reversions that reflect fluctuating immunological dynamics—either loss of pre-existing immunity or positive selection in a newly susceptible populations[30]—in the local communities.

Contemporary sampling and sequencing of RSVB particularly in Africa is still insufficient, compared to HIV-1 viruses, for which the regional spatial ecology has been extensively characterized[50,51]. Paucity of sequence data from neighboring countries limited our inferences on virus importations and movement within the sub-Saharan region. Future studies on local epidemics, transmission and spatial dynamics of the genotype BA could greatly benefit from improved sampling and sequencing in the region. Nonetheless, this study highlights the value of longitudinal sampling within a discrete location and of sequencing sufficiently over a long time to characterize RSV population dynamics. Further studies are required to determine the importance of variant-specific genetic differences in immune response and virus transmission. While the G gene is the fastest evolving RSV gene and has been extensively useful in RSV molecular epidemiology, studying the diversity of the rest of the RSV genome could allow for a deeper understanding of the long-term evolutionary dynamics of the RSV genotype BA.

Methods

Study samples and epidemic patterns. This study was part of longitudinal surveillance study established to understand the epidemiology and disease burden of RSV associated pneumonia[21]. The respiratory samples were collected into the universal transport media (UTM; Copan Diagnostics, USA). We used sequence data from RSV positive samples (nasal washes, aspirates and swabs) collected between January 2003 and August 2017. A majority of the samples were from children < 5 years admitted to Kilifi County Hospital (KCH) (2004–2017) with lower respiratory tract illness[20,21]. The 2003 samples were from the Kilifi RSV birth cohort acute respiratory infection cases[52]. All sample sets arise from the same catchment population and were processed similarly in the laboratory as detailed below.

Informed consent was sought from parents or guardians, and the study protocols were reviewed and approved by the Scientific and Ethical Review Unit, KEMRI, Kenya. All methods were carried out in accordance with relevant guidelines and regulations.

RSV epidemics in Kilifi typically start in November and run through to May with sporadic (inter-epidemic) infections occurring in June to October[21]. However, there is year-to-year variation and therefore we took a pragmatic approach and defined an epidemic as running from October of one year through to September of the following year. The dominant RSV group within a given epidemic was associated with ≥ 60% of the cases; otherwise, the groups were considered co-dominant.

Viral detection, amplification and sequencing. RSV detection was performed by immunofluorescence antibody test (IFAT; RSV DFA kit, Light Diagnostics, UK). Beginning 2007, a real-time RT-PCR assay with primer/probe sets targeting the highly conserved nucleoprotein (N) gene was used to discriminate RSVA and RSVB for all respiratory samples collected during the surveillance period as previously described[24,53,54]. In 2015, the real-time RT-PCR assay for RSVB was updated following failed detection of drifted variants[55]. For genotyping purposes, a 900 base pair (bp) fragment of the G gene was amplified from purified viral nucleic acids of RSVB positive samples and sequenced in a nested PCR reaction as previously described[18,53]. Sequence assembly was done using Sequencher v5.0 (Gene Codes Corp., USA). The RSVB genotype BA viruses were identified following sequence alignments and phylogenetic analysis of the assembled sequences.

Global data set. In addition to the Kilifi sequences, we retrieved RSVB G gene sequences collected between 2012 and 2017 from GenBank. We excluded sequences were shorter than 700 nucleotides, non-BA genotype sequences and those without sampling date or location information. Sequences were aligned using MAFFT 7.222 and alignments manually edited in Aliview[56].

Phylogenetic analyses. Model selection (ModelFinder) and Maximum likelihood (ML) phylogenetic trees were estimated with IQTREE 1.6[57,58]. Node support values were estimated using bootstrap resampling (1000 replicates). Temporal signal in the data was examined using TempEst[59]. Phylogenetic relationships and viral demographic histories were inferred using BEAST 1.10[60]. We used a GTR substitution model with discrete gamma distributed rate variation among sites and uncorrelated relaxed molecular clock with branch rates drawn form a lognormal distribution to account for evolutionary rate variation among lineages. A Skyride demographic prior with time-aware smoothing[61] was selected and a CTMC rate reference prior was specified for the mean clock rate. Chain length of MCMC sampling was 300 million generations, sampling every 10,000. Stationarity and mixing were examined using Tracer 1.7 and maximum clade credibility (MCC) trees summarized using TreeAnnotator.

A variant was defined as a virus or group of viruses with 4 nucleotide differences compared to other viruses and/or falling into a distinct cluster with at least 60% bootstrap support[24].

Selection analysis. We tested the G gene data for evidence of natural selection using methods implemented in the Datamonkey 2.0 webserver[62]. Mean non-synonymous to synonymous rate ratio (d_N/d_S) was estimated using the SLAC method. Pervasive selection analyses were performed using SLAC and FEL methods. Episodic (frequently transient) selection was evaluated using MEME, and the FUBAR method was used to evaluate the difference between nonsynonymous and synonymous rates per codon site. Significance of SLAC, FEL and MEME results used a *P*-value cutoff of 0.05, and FUBAR results were assessed posterior probability of 0.9.

N-linked glycosylation sites. The N-Glycosite web tool (http://www.hiv.lanl.gov/content/sequence/GLYCOSITE/glycosite.html) was used to identify putative N-glycosylation sites (amino acid configuration N-x-S/T, where x is not Proline (P)).

Data availability

The Kilifi G gene sequences analysed here are available in GenBank (accession numbers KP862065-KP862529, KX775722-KX775849 and MH742792-MH742925). Other dataset used in the analysis are found in https://doi.org/10.7910/DVN/XJUA0Z.

Received: 7 August 2020; Accepted: 20 November 2020
Published online: 03 December 2020

References

1. Nair, H. *et al.* Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis. *Lancet* **375**, 1545–1555. https://doi.org/10.1016/s0140-6736(10)60206-1 (2010).
2. Scheltema, N. M. *et al.* Global respiratory syncytial virus-associated mortality in young children (RSV GOLD): a retrospective case series. *Lancet Glob. Health* **5**, e984–e991. https://doi.org/10.1016/s2214-109x(17)30344-3 (2017).
3. Shi, T. *et al.* Global, regional, and national disease burden estimates of acute lower respiratory infections due to respiratory syncytial virus in young children in 2015: a systematic review and modelling study. *Lancet (London, England)* **390**, 946–958. https://doi.org/10.1016/S0140-6736(17)30938-8 (2017).
4. Nyiro, J. U. *et al.* Surveillance of respiratory viruses in the outpatient setting in rural coastal Kenya: baseline epidemiological observations. *Wellcome Open Res.* **3**, 89. https://doi.org/10.12688/wellcomeopenres.14662.1 (2018).
5. Nam, H. H. & Ison, M. G. Respiratory syncytial virus infection in adults. *BMJ* **366**, l5021. https://doi.org/10.1136/bmj.l5021 (2019).
6. Obando-Pacheco, P. *et al.* Respiratory syncytial virus seasonality: a global overview. *J. Infect. Dis.* **217**, 1356–1364. https://doi.org/10.1093/infdis/jiy056 (2018).
7. Neuzil, K. M. Progress toward a respiratory syncytial virus vaccine. *Clin. Vaccine Immunol.* **23**, 186–188 (2016).
8. Higgins, D., Trujillo, C. & Keech, C. Advances in RSV vaccine research and development—a global agenda. *Vaccine* **34**, 2870–2875. https://doi.org/10.1016/j.vaccine.2016.03.109 (2016).
9. McLellan, J. S., Ray, W. C. & Peeples, M. E. Structure and function of respiratory syncytial virus surface glycoproteins. *Curr. Top. Microbiol. Immunol.* **372**, 83–104. https://doi.org/10.1007/978-3-642-38919-1_4 (2013).
10. Cane, P. A. Molecular epidemiology of respiratory syncytial virus. *Rev. Med. Virol.* **11**, 103–116 (2001).
11. Johnson, P. R., Spriggs, M. K., Olmsted, R. A. & Collins, P. L. The G glycoprotein of human respiratory syncytial viruses of subgroups A and B: extensive sequence divergence between antigenically related proteins. *Proc. Natl. Acad. Sci. USA* **84**, 5625–5629 (1987).
12. Katzov-Eckert, H., Botosso, V. F., Neto, E. A., Zanotto, P. M. & Consortium V. Phylodynamics and dispersal of HRSV entails its permanence in the general population in between yearly outbreaks in children. *PLoS ONE* **7**, e41953. https://doi.org/10.1371/journal.pone.0041953 (2012).
13. Zlateva, K. T., Lemey, P., Moes, E., Vandamme, A. M. & Van Ranst, M. Genetic variability and molecular evolution of the human respiratory syncytial virus subgroup B attachment G protein. *J. Virol.* **79**, 9157–9167. https://doi.org/10.1128/jvi.79.14.9157-9167.2005 (2005).
14. Trento, A. *et al.* Major changes in the G protein of human respiratory syncytial virus isolates introduced by a duplication of 60 nucleotides. *J. Gen. Virol.* **84**, 3115–3120. https://doi.org/10.1099/vir.0.19357-0 (2003).
15. Trento, A. *et al.* Ten years of global evolution of the human respiratory syncytial virus BA genotype with a 60-nucleotide duplication in the G protein gene. *J. Virol.* **84**, 7500–7512. https://doi.org/10.1128/JVI.00345-10 (2010).
16. Munywoki, P. K. *et al.* Frequent asymptomatic respiratory syncytial virus infections during an epidemic in a rural Kenyan household cohort. *J. Infect. Dis.* **212**, 1711–1718. https://doi.org/10.1093/infdis/jiv263 (2015).
17. Munywoki, P. K. *et al.* Severe lower respiratory tract infection in early infancy and pneumonia hospitalizations among children Kenya. *Emerg. Infect. Dis.* **19**, 223–229. https://doi.org/10.3201/eid1902.120940 (2013).
18. Agoti, C. N. *et al.* Genetic relatedness of infecting and reinfecting respiratory syncytial virus strains identified in a birth cohort from rural Kenya. *J. Infect. Dis.* **206**, 1532–1541. https://doi.org/10.1093/infdis/jis570 (2012).
19. Berkley, J. A. *et al.* Viral etiology of severe pneumonia among Kenyan infants and children. *JAMA* **303**, 2051–2057. https://doi.org/10.1001/jama.2010.675 (2010).
20. Hammitt, L. L. *et al.* A preliminary study of pneumonia etiology among hospitalized children in Kenya. *Clin. Infect. Dis.* **54**(Suppl 2), S190-199. https://doi.org/10.1093/cid/cir1071 (2012).
21. Nokes, D. J. *et al.* Incidence and severity of respiratory syncytial virus pneumonia in rural Kenyan children identified through hospital surveillance. *Clin. Infect. Dis.* **49**, 1341–1349. https://doi.org/10.1086/606055 (2009).
22. Nokes, D. J. *et al.* Respiratory syncytial virus infection and disease in infants and young children observed from birth in Kilifi District Kenya. *Clin. Infect. Dis.* **46**, 50–57. https://doi.org/10.1086/524019 (2008).
23. Okiro, E. A., Ngama, M., Bett, A. & Nokes, D. J. The incidence and clinical burden of respiratory syncytial virus disease identified through hospital outpatient presentations in Kenyan children. *PLoS ONE* **7**, e52520. https://doi.org/10.1371/journal.pone.0052520 (2012).
24. Agoti, C. N. *et al.* Successive respiratory syncytial virus epidemics in local populations arise from multiple variant introductions, providing insights into virus persistence. *J. Virol.* **89**, 11630–11642. https://doi.org/10.1128/JVI.01972-15 (2015).
25. Munywoki, P. K. *et al.* The source of respiratory syncytial virus infection in infants: a household cohort study in rural Kenya. *J. Infect. Dis.* **209**, 1685–1692. https://doi.org/10.1093/infdis/jit828 (2014).
26. Agoti, C. N. *et al.* Local evolutionary patterns of human respiratory syncytial virus derived from whole-genome sequencing. *J. Virol.* **89**, 3444–3454. https://doi.org/10.1128/jvi.03391-14 (2015).
27. Grubaugh, N. D. *et al.* Tracking virus outbreaks in the twenty-first century. *Nat. Microbiol.* **4**, 10–19. https://doi.org/10.1038/s41564-018-0296-2 (2019).
28. Wohl, S., Schaffner, S. F. & Sabeti, P. C. Genomic analysis of viral outbreaks. *Annu. Rev. Virol.* **3**, 173–195. https://doi.org/10.1146/annurev-virology-110615-035747 (2016).
29. Bennett, S. N. *et al.* Epidemic dynamics revealed in dengue evolution. *Mol. Biol. Evol.* **27**, 811–818. https://doi.org/10.1093/molbev/msp285 (2010).
30. Botosso, V. F. *et al.* Positive selection results in frequent reversible amino acid replacements in the G protein gene of human respiratory syncytial virus. *PLoS Pathog.* **5**, e1000254. https://doi.org/10.1371/journal.ppat.1000254 (2009).
31. Melero, J. A., Mas, V. & McLellan, J. S. Structural, antigenic and immunogenic features of respiratory syncytial virus glycoproteins relevant for vaccine development. *Vaccine* **35**, 461–468. https://doi.org/10.1016/j.vaccine.2016.09.045 (2017).
32. Langedijk, J. P., Schaaper, W. M., Meloen, R. H. & van Oirschot, J. T. Proposed three-dimensional model for the attachment protein G of respiratory syncytial virus. *J. Gen. Virol.* **77**(Pt 6), 1249–1257. https://doi.org/10.1099/0022-1317-77-6-1249 (1996).
33. Fedechkin, S. O., George, N. L., Wolff, J. T., Kauvar, L. M. & DuBois, R. M. Structures of respiratory syncytial virus G antigen bound to broadly neutralizing antibodies. *Sci. Immunol.* https://doi.org/10.1126/sciimmunol.aar3534 (2018).
34. Collarini, E. J. *et al.* Potent high-affinity antibodies for treatment and prophylaxis of respiratory syncytial virus derived from B cells of infected patients. *J. Immunol.* **183**, 6338–6345. https://doi.org/10.4049/jimmunol.0901373 (2009).

35. Jones, H. G. *et al.* Structural basis for recognition of the central conserved region of RSV G by neutralizing human antibodies. *PLoS Pathog.* **14**, e1006935. https://doi.org/10.1371/journal.ppat.1006935 (2018).
36. Murray, G. G. *et al.* The effect of genetic structure on molecular dating and tests for temporal signal. *Methods Ecol. Evol.* **7**, 80–89. https://doi.org/10.1111/2041-210X.12466 (2016).
37. Melero, J. A. & Moore, M. L. Influence of respiratory syncytial virus strain differences on pathogenesis and immunity. *Curr. Top. Microbiol. Immunol.* **372**, 59–82. https://doi.org/10.1007/978-3-642-38919-1_3 (2013).
38. Peret, T. C. *et al.* Circulation patterns of group A and B human respiratory syncytial virus genotypes in 5 communities in North America. *J. Infect. Dis.* **181**, 1891–1896. https://doi.org/10.1086/315508 (2000).
39. Trovao, N. S. *et al.* Molecular characterization of respiratory syncytial viruses circulating in a paediatric cohort in Amman Jordan. *Microb. Genom.* https://doi.org/10.1099/mgen.0.000292 (2019).
40. Gaunt, E. R. *et al.* Molecular epidemiology and evolution of human respiratory syncytial virus and human metapneumovirus. *PLoS ONE* **6**, e17427. https://doi.org/10.1371/journal.pone.0017427 (2011).
41. Hotard, A. L., Laikhter, E., Brooks, K., Hartert, T. V. & Moore, M. L. Functional analysis of the 60-nucleotide duplication in the respiratory syncytial virus buenos aires strain attachment glycoprotein. *J. Virol.* **89**, 8258–8266. https://doi.org/10.1128/JVI.01045-15 (2015).
42. Trento, A. *et al.* Conservation of G-protein epitopes in respiratory syncytial virus (Group A) despite broad genetic diversity: is antibody selection involved in virus evolution?. *J. Virol.* **89**, 7776–7785. https://doi.org/10.1128/JVI.00467-15 (2015).
43. Delport, W., Scheffler, K. & Seoighe, C. Frequent toggling between alternative amino acids is driven by selection in HIV-1. *PLoS Pathog.* **4**, e1000242. https://doi.org/10.1371/journal.ppat.1000242 (2008).
44. Cornelissen, M., Hogervorst, E., Zorgdrager, F., Hartman, S. & Goudsmit, J. Maintenance of syncytium-inducing phenotype of HIV type 1 is associated with positively charged residues in the HIV type 1 gp120 V2 domain without fixed positions, elongation, or relocated N-linked glycosylation sites. *AIDS Res. Hum. Retroviruses* **11**, 1169–1175. https://doi.org/10.1089/aid.1995.11.1169 (1995).
45. Gambaryan, A. S., Matrosovich, M. N., Bender, C. A. & Kilbourne, E. D. Differences in the biological phenotype of low-yielding (L) and high-yielding (H) variants of swine influenza virus A/NJ/11/76 are associated with their different receptor-binding activity. *Virology* **247**, 223–231. https://doi.org/10.1006/viro.1998.9274 (1998).
46. Brault, A. C., Powers, A. M., Holmes, E. C., Woelk, C. H. & Weaver, S. C. Positively charged amino acid substitutions in the e2 envelope glycoprotein are associated with the emergence of venezuelan equine encephalitis virus. *J. Virol.* **76**, 1718–1730. https://doi.org/10.1128/jvi.76.4.1718-1730.2002 (2002).
47. Hensley, S. E. *et al.* Hemagglutinin receptor binding avidity drives influenza A virus antigenic drift. *Science* **326**, 734–736. https://doi.org/10.1126/science.1178258 (2009).
48. Esposito, S. *et al.* Characteristics and their clinical relevance of respiratory syncytial virus types and genotypes circulating in Northern Italy in five consecutive winter seasons. *PLoS ONE* **10**, e0129369. https://doi.org/10.1371/journal.pone.0129369 (2015).
49. Tan, L. *et al.* The comparative genomics of human respiratory syncytial virus subgroups A and B: genetic variability and molecular evolutionary dynamics. *J. Virol.* **87**, 8213–8226. https://doi.org/10.1128/JVI.03278-12 (2013).
50. Wilkinson, E. *et al.* Origin, imports and exports of HIV-1 subtype C in South Africa: a historical perspective. *Infect. Genet. Evol.* **46**, 200–208. https://doi.org/10.1016/j.meegid.2016.07.008 (2016).
51. Raghwani, J. *et al.* Evolution of HIV-1 within untreated individuals and at the population scale in Uganda. *PLoS Pathog.* **14**, e1007167. https://doi.org/10.1371/journal.ppat.1007167 (2018).
52. Nokes, D. J. *et al.* Respiratory syncytial virus epidemiology in a birth cohort from Kilifi district, Kenya: infection during the first year of life. *J. Infect. Dis.* **190**, 1828–1832. https://doi.org/10.1086/425040 (2004).
53. Otieno, J. R. *et al.* Molecular evolutionary dynamics of respiratory syncytial virus group A in recurrent epidemics in coastal Kenya. *J. Virol.* **90**, 4990–5002. https://doi.org/10.1128/jvi.03105-15 (2016).
54. Hammitt, L. L. *et al.* Added value of an oropharyngeal swab in detection of viruses in children hospitalized with lower respiratory tract infection. *J. Clin. Microbiol.* **49**, 2318–2320. https://doi.org/10.1128/jcm.02605-10 (2011).
55. Kamau, E. *et al.* Recent sequence variation in probe binding site affected detection of respiratory syncytial virus group B by real-time RT-PCR. *J. Clin. Virol.* **88**, 21–25. https://doi.org/10.1016/j.jcv.2016.12.011 (2017).
56. Larsson, A. AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics (Oxford, England)* **30**, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531 (2014).
57. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589. https://doi.org/10.1038/nmeth.4285 (2017).
58. Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274. https://doi.org/10.1093/molbev/msu300 (2015).
59. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007. https://doi.org/10.1093/ve/vew007 (2016).
60. Suchard, M. A. *et al.* Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* **4**, vey016. https://doi.org/10.1093/ve/vey016 (2018).
61. Minin, V. N., Bloomquist, E. W. & Suchard, M. A. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol.* **25**, 1459–1471. https://doi.org/10.1093/molbev/msn090 (2008).
62. Delport, W., Poon, A. F., Frost, S. D. & Kosakovsky Pond, S. L. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics* **26**, 2455–2457. https://doi.org/10.1093/bioinformatics/btq429 (2010).

Acknowledgements

We are thankful to the laboratory, field, and clinical staff for sample and clinical data collection and laboratory testing. We are also thankful to the guardians and parents of the children who participated in this study. This paper is published with the permission of the Director of KEMRI. This study was funded by The Wellcome Trust (102975, 203077). Dr Charles Agoti is supported by the Initiative to Develop African Research Leaders (IDeAL) through the DELTAS Africa Initiative (DEL-15-003). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [107769/Z/10/Z] and the UK government. The views expressed in this publication are those of the authors and not necessarily those of AAS, NEPAD Agency, Wellcome Trust or the UK government.

Author contributions

E.K. performed sequence analysis and wrote the manuscript drafts, J.R.O. was involved in data collection and sequence analysis, C.L. and A.M. prepared samples for sequencing, N.M. was involved in patient data collection

and statistical analysis, D.J.N. and C.N.A. conceived and designed the research. All authors read and reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-78234-0.

Correspondence and requests for materials should be addressed to E.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.